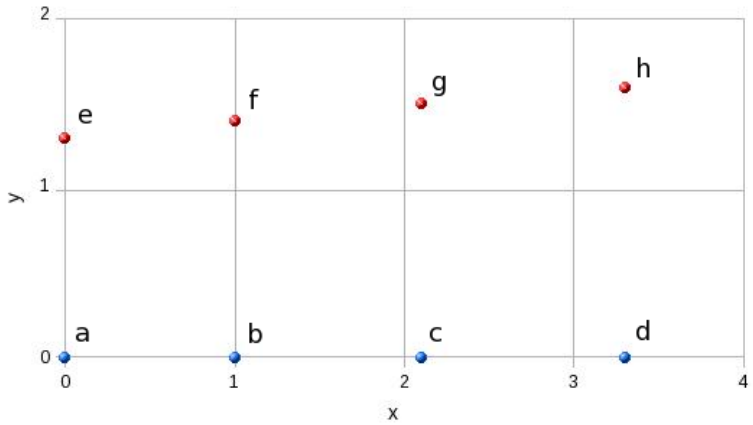
M33SAD – průběžný test znalostí

Shlukování, vyhledávání častých vzorů

1 Shlukování

Mějme množinu 8 příkladů $\mathcal{X} = \{a = (0,0), b = (1,0), c = (2.1,0), d = (3.3,0), e = (0,1.3), f = (1,1.4), g = (2.1,1.5), h = (3.3,1.6)\}$. Předpokládejme, že skutečný rozklad příkladů (gold standard) je $G = \{\{a,b,c,d\},\{e,f,g,h\}\}$. Tento rozklad nemá shlukovací algoritmus k dispozici, bude sloužit pouze k jeho hodnocení. Dopředu známý je počet shluků $k = 2$. Zodpovězte následující otázky:



(a) (2 body) Předpokládejme, že data jsou generována směsí dvou rozdělení. Parametry rozdělení 1 a 2 odhaduje EM algoritmus pomocí skryté proměnné Z a modelu θ. V aktuálním kroku t jsou váhy jednotlivých rozdělení $\alpha_1 = 0.9$ a $\alpha_2 = 0.1$. Vypočtěte $P(Z_i = j|X_i;\theta^{(t)})$, tedy pravděpodobnosti příslušnosti k rozdělení vzhledem k aktualnímu modelu. Hustoty v bodech (a,b,c,d,e,f,g,h) jsou v tomto kroku při modelu $\theta^{(t)}$ pro první a druhé rozdělení $(0.08, 0.02, 0, 0, 0.12, 0.32, 0.11, 0)$ a $(0, 0.09, 0.23, 0.08, 0, 0, 0.03, 0.06)$ respektive.

$$P(Z_i = j|X_i;\theta^{(t)}) = \frac{\alpha_j f_{i,j}}{\alpha_1 f_{i,1} + \alpha_2 f_{i,2}}$$

```
t1=0.9;t2=0.1
t1*0.08/(t1*0.08+t2*0) #a
t1*0.02/(t1*0.02+t2*0.09) #b
t1*0.0/(t1*0.0+t2*0.23) #c
t1*0.0/(t1*0.0+t2*0.08) #d
t1*0.12/(t1*0.12+t2*0) #e
t1*0.32/(t1*0.32+t2*0) #f
```

t1*0.11/(t1*0.11+t2*0.03) #g
t1*0.0/(t1*0.0+t2*0.06) #h

Výsledek - příslušnost—spravná třída
1 - 1 —1
0.666- 1 —1
0 - 2 —1
0 - 2 —1
1 - 1 —2
1 - 1 —2
0.97 - 1 —2
0 - 2 —2

(b) (1 bod) Spočítejte ryzost rozkladu na shluky z bodu a, uvažujte "ostrou" příslušnost ke shlukům na základě kritéria maximální pravděpodobnosti.

$$purity = \frac{3+2}{8} = 0.625$$

$$G = \{\{1,1,2,2,2\},\{1,1,2\}\}$$

$$RI = \frac{\binom{3}{2} + \binom{2}{2} + 2 \cdot 1 + 3 \cdot 2}{\binom{8}{2}} = \frac{12}{28} = 0.429$$

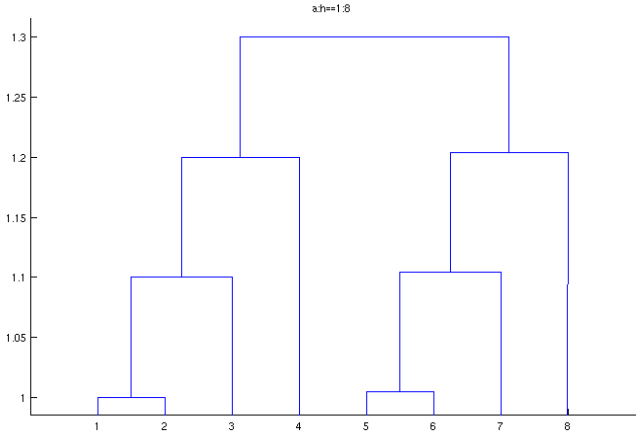
(c) (1 bod) Matematicky formulujte kritérium, které je optimalizováno algoritmem *k-středů*.

Globální kritérium homogenity:

$$W(k) = \arg\min_{\Omega} \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, \mu_i)$$

Kde Ω je rozklad.

(d) (1 bod) Zakreslete dendrogram generovaný algoritmem hierarchického shlukování (*single linkage*) na datech z obrázku.



(e) (2 body) Popište algoritmus divizivního hierarchického shlukovaní pomocí *k-středů* ($k = 2$) a zakreslete první dva kroky (inicializujte výpočet body $\{f, g\}$, případné následující instance *k-středů* inicializujte dle svého uvážení).

Rekurzivně voláme k-means:

$$G = |\{a, b, e, d\}, \{c, d, g, h\}| e.g. \{a, b\}, \{e, d\}, \{c, d\}, \{g, h\}|$$

2 Časté podgrafy

V následující úloze řekneme, že nějaký graf g je *následníkem* jiného grafu g' (g' je *předchůdcem* g), když g obsahuje všechny hrany g' a právě jednu hranu navíc.
Mějme množinu tří grafů z obrázku. Vrcholy jsou anotované značkami A, nebo B, hrany mají identické značky.

(a) (3 body) Nakreslete strom takový, že každý podgraf alespoň jednoho grafu ze zadání je v tomto stromě právě jednou a hrany stromu vedou vždy z nějakého podgrafu do jeho následníka.

(b) (2 body) Uvažujte minimální podporu $s_{min} = 2$, vyznačte všechny uzavřené a maximální podgrafy ve výše zkonstruovaném stromu. Pro přehlednost si můžete strom doplnit (např. čárkovanými hranami) na orientovaný acyklický graf, v němž do každého podgrafu vedou hrany ze všech jeho předchůdců.

- Maximální $\{B - A\}, \{A - A\}$
- Uzavřený $\{A\}, \{B - A\}, \{A - A\}$

(c) (1 bod) V kontextu vyhledávání podgrafů popište monotónní (resp. antimonotónní) vlastnost, na které je založen algoritmus APRIORI.

Podgraf, který není častý, nemůže mít častý nadgraf.